Nucleic Acids Research, 2000, Vol. 28, No. 1 27–30

KEGG: Kyoto Encyclopedia of Genes and Genomes

Minoru Kanehisa* and Susumu Goto

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Received September 29, 1999; Accepted October 4, 1999

ABSTRACT

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The genomic information is stored in the GENES database, which is a collection of gene catalogs for all the completely sequenced genomes and some partial genomes with up-to-date annotation of gene functions. The higher order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. The PATHWAY database is supplemented by a set of ortholog group tables for the information about conserved subpathways (pathway motifs), which are often encoded by positionally coupled genes on the chromosome and which are especially useful in predicting gene functions. A third database in KEGG is LIGAND for the information about chemical compounds, enzyme molecules and enzymatic reactions. KEGG provides Java graphics tools for browsing genome maps, comparing two genome maps and manipulating expression maps, as well as computational tools for sequence comparison, graph comparison and path computation. The KEGG databases are daily updated and made freely available (http://www.genome.ad.jp/kegg/).

INTRODUCTION

While the genome sequencing projects rapidly determine gene catalogs for an increasing number of organisms, functional annotation of individual genes is still largely incomplete. KEGG (Kyoto Encyclopedia of Genes and Genomes) is an effort to link genomic information with higher order functional information by computerizing current knowledge on cellular processes and by standardizing gene annotations. Generally speaking, the biological function of the living cell is a result of many interacting molecules; it cannot be attributed to just a single gene or a single molecule (1). The functional assignment in KEGG is a process of linking a set of genes in the genome with a network of interacting molecules in the cell, such as a pathway or a complex, representing a higher order biological function.

The KEGG project was initiated in May 1995 under the Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan (2). All the data in KEGG and associated software tools are made available as part of the Japanese GenomeNet service (3). KEGG consists of three databases: PATHWAY for representation of higher order functions in terms of the network of interacting molecules, GENES for the collection of gene catalogs for all the completely sequenced genomes and some partial genomes, and LIGAND (4) for the collection of chemical compounds in the cell, enzyme molecules and enzymatic reactions. The overall architecture of the KEGG system is basically the same as previously reported (5). The user may enter the KEGG system top-down starting from the pathway (functional) information or bottom-up starting from the genomic information at the KEGG table of contents page (http://www.genome.ad.jp/kegg/kegg2.html).

GENOMIC INFORMATION

GENES database

The current status of the KEGG databases is summarized in Table 1. During the past year, we have made every effort to keep up with the data increase of complete genome sequences, and also with the imminent data explosion of gene expression profiles. The number of GENES entries for just 29 species—human, mouse, *Drosophila*, *Arabidopsis*, *Schizosaccharomyce pombe*, and 24 completely sequenced genomes—totals ~110 000 entries, which is already larger than the number of entries in the well-annotated SWISS-PROT database (6). The GENES database contains the bare minimum information for each gene as shown in Table 2, but it is intended to be a resource containing up-to-date, standardized descriptions of gene functions. GENES also serves as a gateway to a number of other resources containing more detailed information.

We developed various computational tools for the maintenance of the GENES database, especially for extraction of information from GenBank, which is not a trivial task, and for assisting systematic annotation of gene functions. The overall flow of both computerized and manual processes is illustrated in Figure 1. A web-based annotation tool is used, together with other computational tools, to assign EC numbers, to assign ortholog identifiers, to incorporate new experimental evidence from literature, and to annotate predictions based on pathway construction. As described below, the ortholog identifiers will be used as primary keys for automatic matching of genes in the genome and gene products in the pathway.

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp

Table 1. The summary of KEGG release 12.0 (October 1999)

Database	Content
PATHWAY	2706 entries for pathway diagrams constructed from 143 manually drawn diagrams
GENES	110 018 entries in 24 complete genomes and 12 partial genomes
LIGAND	5645 entries in the COMPOUND section
	3705 entries in the ENZYME section
	5207 reactions in the REACTION section
Auxiliary data	**Content**
Ortholog group table	61 tables
Genome map	23 complete genomes and one partial genome
Comparative genome map	23 × 23 complete genome comparisons
Expression map	Four sets of expression maps
Gene catalog	53 catalogs
Molecular catalog	Eight catalogs
Disease catalog	Three catalogs

Table 2. The data content of the GENES database entry

Field	Content	Links	Data source
ENTRY	Entry identifier (gene accession number)	LinkDB database	GenBank or original database
NAME	Gene names and alternative names		GenBank or original database
DEFINITION	Annotation of gene function	LIGAND/ENZYME database, SWISS-PROT database and PubMed database	GenBank, original database, SWISS-PROT and KEGG
CLASS	Classification of genes according to the KEGG pathways	KEGG/PATHWAY database	KEGG
POSITION	Chromosomal position	KEGG/GENOME map	GenBank
DBLINKS	Outside links	Original databases and NCBI Entrez database	
CODON_USAGE	Codon usage		Computed
AASEQ	Amino acid sequence	see footnote[a]	GenBank or original database
NTSEQ	Nucleotide sequence		GenBank or original database

[a]Computational links are available including sequence similarity searches (FASTA and BLAST), motif search (MOTIF), membrane protein predictions (SOSUI and TSEG), and cellular localization site prediction (PSORT).

Gene expression profiles

The backbone retrieval system for the GENES database is the DBGET/LinkDB system (7), but there are additional ways of accessing this database. One is the Java-based genome map browser for graphical manipulation of gene positions on the chromosome. The other is what we call the hierarchical text browser for handling functional hierarchy of gene catalogs. Here we report another Java graphics browser, the expression map browser, for analysis of gene expression profiles obtained by cDNA microarray or oligonucleotide array experiments. The vast amount of data generated by such functional genomics experiments are likely to contain valuable information, which will supplement genomic sequence information toward understanding higher biological functions of the cell. A preliminary version of the expression map browser is linked to both the KEGG pathway data and the genome map data, so that the user may examine if, for example, a group of co-regulated genes are also correlated in the pathway or are encoded in a cluster of genes on the chromosome.

PATHWAY INFORMATION

PATHWAY database

Currently the best organized part of the KEGG/PATHWAY database is metabolism, which is represented by ~90 graphical diagrams for the reference metabolic pathways. Each reference pathway can be viewed as a network of enzymes or a network of EC numbers. Once enzyme genes are identified in the genome based on sequence similarity and positional correlation of genes, and the EC numbers are properly assigned, organism-specific pathways can be constructed computationally by correlating genes in the genome with gene products (enzymes)

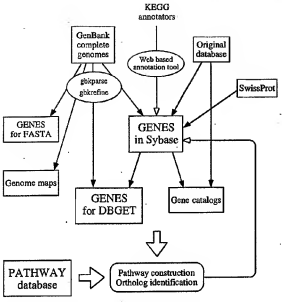

Figure 1. Procedures used to organize and annotate the GENES database.

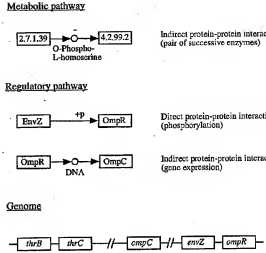

Figure 2. The generalized protein–protein interaction includes an indirect protein–protein interaction by two successive enzymes, a direct protein–protein interaction, and another indirect protein–protein interaction by gene expression. The nodes of the generalized protein–protein interaction network are gene products, which can be directly correlated with genes in the genome.

in the reference pathways according to the matching EC numbers. We are trying to extend this mechanism to include various regulatory pathways, such as signal transduction, cell cycle and apoptosis. There are, however, two major problems in automating the construction of regulatory pathways.

Because the metabolic pathway, especially for intermediary metabolism, is well conserved among most organisms from mammals to bacteria, it is possible to manually draw one reference pathway and then to computationally generate many organism-specific pathways. In contrast, the regulatory pathways are far more divergent and are difficult to combine into common reference pathway diagrams. Thus, we basically draw a pathway diagram separately for each organism. At the same time, we are trying to identify groups of organisms that share common pathways or assemblies and whose diagrams may be combined. Examples include one common apoptosis pathway diagram for human and mouse, three ribosome assembly diagrams separately for bacteria, archaea and eukaryotes.

The other related problem is the absence of proper identifiers for functions in the regulatory pathways. The EC numbers in the metabolic pathways play roles as identifiers of the nodes (enzymes) and also as keys for linking with the genomic information. We are preparing for the introduction of the ortholog identifiers to extend such capabilities of the EC numbers. The ortholog identifiers will be used to identify nodes (proteins) in the regulatory pathways and also to link with the genomic information. In addition, the ortholog identifiers will replace the EC numbers in the metabolic pathways in order to distinguish multiple genes that match one EC number, for example, different subunits of an enzyme complex or different genes expressed under different conditions.

Ortholog group tables

Orthologs are identified in KEGG not only by sequence similarity of individual genes but also by examining if all constituent members are found for a functional group, such as a conserved subpathway or a molecular complex. The KEGG ortholog group table is a representation of three features: whether an organism contains a complete set of genes that constitutes a functional group, whether those genes are physically coupled on the chromosome, and what are orthologous genes among different organisms. Currently there are 61 ortholog group tables, which contain, for example, a gene cluster in the genome coding for a functionally related enzyme cluster in the metabolic pathway. In KEGG such correlated clusters are first detected by a heuristic graph comparison algorithm, and then manually edited and compiled into the ortholog group tables. There are two types of graph comparisons that we use: genome–pathway and genome–genome comparisons (1). An ortholog group table is a composite of such pairwise comparisons, representing a conserved portion of the pathway, or what we call a pathway motif.

Generalized protein–protein interaction

The KEGG pathway representation focuses on the network of gene products, mostly proteins but including functional RNAs. As illustrated in Figure 2, the metabolic pathway is a network of indirect protein–protein interactions, which is actually a network of enzyme–enzyme relations. In contrast, the regulatory pathway often consists of direct protein–protein interactions, such as binding and phosphorylation, and another class of indirect protein–protein interactions, which are relations of transcription factors and transcribed gene products via gene expressions. The generalized protein–protein interaction network that includes these three types of interactions is an

abstract network, but it is especially useful to link with genomic information because the nodes (gene products) of this network can be directly correlated with the nodes (genes) in the genome. With this concept of generalized protein–protein interaction network, we are expanding the collection of manually drawn reference pathway diagrams.

AVAILABILITY

All the data in KEGG and associated analysis tools are provided as part of the Japanese GenomeNet service (3) at http://www.genome.ad.jp/

The Internet version of KEGG in GenomeNet can be accessed at the following address: http://www.genome.ad.jp/kegg/

For strictly academic research purposes at academic institutions the KEGG mirror server package may be installed. The package, which also includes a minimal set of DBGET/LinkDB, can be obtained from the KEGG anonymous FTP site: ftp://kegg.genome.ad.jp/

The mirror package runs on a Solaris, IRIX or Linux machine. The individual databases PATHWAY, GENES and LIGAND can also be mirrored or obtained by anonymous FTP.

ACKNOWLEDGEMENTS

We thank the present and past KEGG project members for their excellent work. We also thank Kotaro Shiraishi for developing the KEGG annotation tool and other useful programs. This work was supported by a Grant-in-Aid for Scientific Research on the Priority Area 'Genome Science' from the Ministry of Education, Science, Sports and Culture of Japan. The computational resource was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

REFERENCES

1. Kanehisa,M. (1999) *Post-Genome Informatics*. Oxford University Press, Oxford, UK.
2. Kanehisa,M. (1997) *Trends Genet.*, 13, 375–376.
3. Kanehisa,M. (1997) *Trends Biochem. Sci.*, 22, 442–444.
4. Goto,S., Nishioka,T. and Kanehisa,M. (2000) *Nucleic Acids Res.*, 28, 380–382.
5. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) *Nucleic Acids Res.*, 27, 29–34.
6. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, 27, 49–54. See also this issue: *Nucleic Acids Res.* (2000) 28, 45–48.
7. Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) *Pac. Symp. Biocomput. 1998*, 683–694.